Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predicting E-commerce Revenue using Machine Learning and User Behavior Analysis

Zahraa Hussein Ameen Aldahlaki

Al-Nukhba university college

ARTICLE INFO

Received: 24 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

The digital economy, and in particular, the e-commerce sector, has experienced significant growth in the last few decades [1]. The ability to accurately predict revenue in this burgeoning sphere is crucial for businesses to maximize profitability and to maintain a competitive advantage [2]. However, traditional forecasting methods, primarily based on historical data and statistical techniques, often fail to accurately anticipate the dynamic and nonlinear nature of e-commerce revenue. These techniques, such as time-series analysis, linear regression, and moving average models, often rely on the assumption of linearity and stationarity, which may not hold true in the unpredictable landscape of e-commerce, characterized by rapidly changing consumer behavior, market trends, and competitive dynamics [3].

Keywords: consumer, forecasting, significant, linearity

Introduction

According to Forbes [4], e-commerce sales in the US reached $870 billion in 2021, a 14.2% increase over 2020 and a 50.5% increase over 2019. E-commerce accounted for 13.2% of all retail sales in the US in 2021. Statista reports that retail e-commerce sales worldwide amounted to approximately $5.2 trillion in 2021, and this figure is forecast to grow by 56% over the next few years, reaching about $8.1 trillion by 2026 [5]. E-commerce as a percentage of total retail sales worldwide was 20.4% in 2021 and is expected to reach 25.4% by 2026. Popupsmart reveals that the fastest-growing retail e-commerce market by compound annual growth rate (CAGR) from 2023 to 2027 is Brazil with 17.3%, followed by India with 15.8%, and China with 14.9% [6].

In the face of these limitations, machine learning (ML) has emerged as a powerful tool for predicting e-commerce revenue, offering the potential for superior accuracy and adaptability. ML algorithms, including linear regression models, decision trees, random forests, support vector machines, and neural networks, can process large and complex data sets, capturing non-linear patterns that are often overlooked by traditional methods. These models can leverage not only historical sales data but also other indicators, such as website traffic, product reviews, and social media sentiment, providing a more holistic view of the factors influencing revenue [7] [8].

In the thesis at hand, we propose a novel approach to predicting e-commerce revenue by integrating ML techniques with user behavior analysis. This approach allows us to capture both the broad market trends and the nuanced individual-level variations in consumer behavior, thereby generating more accurate and granular revenue predictions. By combining different sources of information, we create a more complete picture of the consumer, encompassing not just their purchase history, but also their browsing behavior, product preferences, and engagement with the platform. This comprehensive understanding of the user aids in identifying trends, predicting future behaviors, and ultimately, forecasting revenue more accurately. We believe this innovative combination of ML and user behavior analysis paves the way for a new era in e-commerce revenue prediction, marrying advanced technological capabilities with a deep understanding of the customer.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Problem Statement

The problem that the current study intends to address is the critical gap in the effective prediction of e-commerce revenue, an issue that has become increasingly pronounced with the dynamic and complex nature of the online retail landscape. Traditional forecasting models, while providing a foundation, have shown limitations due to their inability to accommodate non-linear, volatile, and multifactorial elements integral to e-commerce. These models, founded primarily on linear statistical methods, often fail to encapsulate the complexity of digital commerce, which involves interactions of a myriad of variables, including but not limited to customer behavior, market competition, product trends, and time-sensitive events such as sales and promotions.

Moreover, the rapid evolution of consumer behavior, driven by changing technologies, market trends, and societal factors, has made predicting e-commerce revenue an increasingly challenging endeavor. Conventional methods may not fully account for these shifts, leading to inaccurate or short-sighted forecasts. The consequences of imprecise predictions can be severe for businesses, including loss of potential profits, misallocation of resources, and impaired strategic planning.

Despite the tremendous potential of ML techniques to address these issues, the application of such models in e-commerce revenue prediction has been relatively underexplored, and existing applications often ignore crucial aspects of the consumer experience. The majority of ML models used to predict e-commerce revenue primarily rely on transactional data and general market trends. However, these models often fail to incorporate the rich tapestry of user behaviors, including browsing patterns, search queries, product reviews, and social media interactions, all of which can offer valuable insights into consumer preferences and future purchasing behavior.

Research Question

- How can the Online Shoppers Purchasing Intention Dataset be effectively analyzed and interpreted to reveal meaningful patterns and relationships that can contribute to predicting e-commerce revenue?
- How can various ML models (random forest, XGBoost, decision tree, SVM, and linear regression) be developed and trained on the dataset to accurately predict e-commerce revenue?
- How do different ML models compare in terms of their accuracy, precision, and robustness in predicting e-commerce revenue? Which model(s) demonstrate the most promising performance?
- How can user behavior analysis be integrated into ML models to better capture both broad market trends and nuanced individual-level variations in consumer behavior?
- How accurate are the revenue predictions generated by the selected ML model(s) when compared to actual revenue figures?
- How can the most effective ML models be implemented in real-world e-commerce settings to improve revenue prediction and facilitate strategic decision-making?
- How does the proposed approach contribute to the existing body of knowledge in the field of ML in e-commerce, and how can it stimulate further research and development in this domain?

Research Objectives

In light of the above, our objectives for this research thesis, 'Predicting E-commerce Revenue using ML and User Behavior Analysis,' are as follows:

- To conduct a comprehensive analysis of the Online Shoppers Purchasing Intention Dataset, identifying key variables and potential relationships that may influence e-commerce revenue. This would encompass user characteristics such as browsing behavior, product preferences, and engagement with the platform.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- To develop and train several ML models, including Random Forest, XGBoost, Decision Tree, SVM, and Linear Regression. These models will be calibrated and tested for their ability to predict e-commerce revenue based on the user behavior data extracted from the dataset.
- To compare the performance of the different ML models in terms of their accuracy, precision, and robustness in predicting e-commerce revenue. Based on these comparisons, we aim to select the model(s) that provide the most reliable and accurate predictions.
- To integrate user behavior analysis into the selected model(s), enabling us to capture both broad market trends and nuanced individual-level variations in consumer behavior. The intention is to generate more accurate and granular revenue predictions than conventional models.
- To evaluate the accuracy of our revenue predictions by comparing them against actual revenue figures. This evaluation will help assess the effectiveness of our approach and identify potential areas for improvement.
- To provide practical guidelines for implementing the most effective ML model(s) in real-world e-commerce settings. We aim to showcase how the integration of ML and user behavior analysis can lead to improved revenue prediction and strategic decision-making in e-commerce businesses.
- To contribute to the burgeoning field of ML in e-commerce by demonstrating a novel application of these models in conjunction with user behavior analysis. Our research aims to encourage further exploration and development of such integrative approaches in revenue prediction and other related areas.

Significance of the Study

The significance of this study is multifold and extends across both practical and theoretical realms. By introducing a novel approach for predicting e-commerce revenue using ML techniques integrated with user behavior analysis, this study fills a critical gap in the current body of research. It aims to advance beyond traditional prediction models, offering more accurate and granular revenue forecasts by incorporating a wide array of consumer behavior data from the "Online Shoppers Purchasing Intention Dataset."

The practical implications for businesses are profound. By enabling more accurate forecasting of e-commerce revenue, businesses can optimize their strategic planning, make more informed decisions, and potentially enhance profitability. The nuanced understanding of consumer behavior that our study offers can inform more personalized marketing strategies and user experiences, which could increase customer satisfaction and loyalty, further boosting revenue.

On a theoretical level, this study pushes the boundaries of what ML models can achieve in the realm of e-commerce. By comparing and evaluating the efficacy of multiple ML models such as Random Forest, XGBoost, Decision Tree, SVM, and Linear Regression, the study contributes significantly to the ongoing dialogue in the fields of ML and predictive analytics. The results of this research could stimulate further innovation and research, ultimately advancing our understanding and application of ML models in predicting e-commerce revenue and beyond.

Related Works

Paper [8] explores the application of ML techniques, specifically deep learning algorithms, in predicting online product sales, an increasingly important area given the exponential growth of online business over the last decade. The research presents a sales prediction model designed for online products, emphasizing the model's adaptability across different types of products. Through a comparative analysis of a fully connected model and a convolutional neural network (CNN), the study proves the superior accuracy and generalization capability of the CNN model. Moreover, the performance advantages of the CNN model are further underscored by comparing it to a non-deep learning model.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The study concludes that an unsupervised pre-trained CNN model is more effective and adaptable in sales forecasting, demonstrating its potential as a valuable tool in understanding user preferences and enhancing online business strategies.

The study [9] explores the role of ML in enhancing sales forecasting within the booming e-commerce industry. E-commerce, characterized by its convenience to customers and its expanding market share, generated $603 billion in the United States alone in 2019. Given this context, the study aims to build ML algorithms capable of forecasting e-commerce sales. The methodology includes a literature review to identify the most effective ML models used in similar studies. The selected models are then built, tested, and evaluated based on their accuracy, error, and overall performance. The goal is to identify the model with the lowest error rate and highest accuracy for sales forecasting, which will then be integrated into the researcher's system. This system aims to provide insights into both current and predicted future sales, offering a valuable tool for e-commerce enterprises.

The research [10] examines the critical role of sales forecasting in the burgeoning e-commerce industry, a sector increasingly crucial to national economies, especially during challenging times such as a global pandemic. The growing competition among e-commerce platforms necessitates superior user service and management, in which accurate sales forecasting is a vital component. Although numerous studies have explored e-commerce sales prediction, the quest for a universally adaptable prediction model remains. This paper evaluates two linear models, three ML models, and two deep learning models for sales forecasting. Interestingly, it finds that while ML and deep learning models don't necessarily enhance forecast accuracy, all models, including linear ones, perform better when they incorporate calendar and price information into the predictions. This finding advances our understanding of effective sales forecasting in the e-commerce realm.

The paper [11] introduces a novel deep neural framework, DSF, for sales forecasting in e-commerce, addressing overlooked factors such as promotion campaigns and the competitive relationship between substitutable products. Traditional forecasting techniques mainly consider sales records alone, often disregarding these complex influences. In DSF, sales forecasting is treated as a sequence-to-sequence learning problem, and a sales residual network is incorporated to model the impact of competition when a promotion campaign is introduced for a product or its alternatives. Extensive experiments are conducted on two real-world datasets from the Taobao e-commerce platform across different domains. The findings indicate that DSF significantly outperforms traditional baselines and cutting-edge deep learning alternatives in terms of forecasting accuracy. In fact, DSF surpasses the current deep learning solution deployed on the Taobao platform, demonstrating its efficacy and potential for broad application in e-commerce sales forecasting.

The paper [12] explores the application of ML algorithms to resource management in cross-border e-commerce, focusing on sales forecasting and inventory optimization within the enterprise resource planning system (ERP). As cross-border e-commerce enterprises expand and their overseas warehousing operating costs rise, ERP systems are tasked with global supply chain warehouse management, coordination, and optimization. Through ML algorithms on the cross-border e-commerce ERP platform, key factors are effectively summarized, and sales record big data is utilized to align forecast values with actual trends. The findings indicate that ML algorithms offer reliable sales predictions and enhance inventory balance efficiency, thus demonstrating their crucial role in optimizing resource management within the context of cross-border e-commerce.

The study [13] focuses on the critical area of market share analysis and sales forecasting, particularly in the context of the e-commerce market, which has been largely overlooked in traditional research. Recognizing the value of predicting sales to guide merchants towards higher profits, the paper utilizes historical e-commerce data to develop sales prediction models. Three types of models are studied: Incentive Auto Regressive Integrated Moving Average (I-ARIMA), Long Short-Term Memory (LSTM), and Artificial Neural Network (ANN). These models cater to different accuracy requirements and data

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

types. The paper compares the strengths and weaknesses of these models across different data sets, providing valuable insights to aid merchants in shaping their marketing strategies.

Methodology

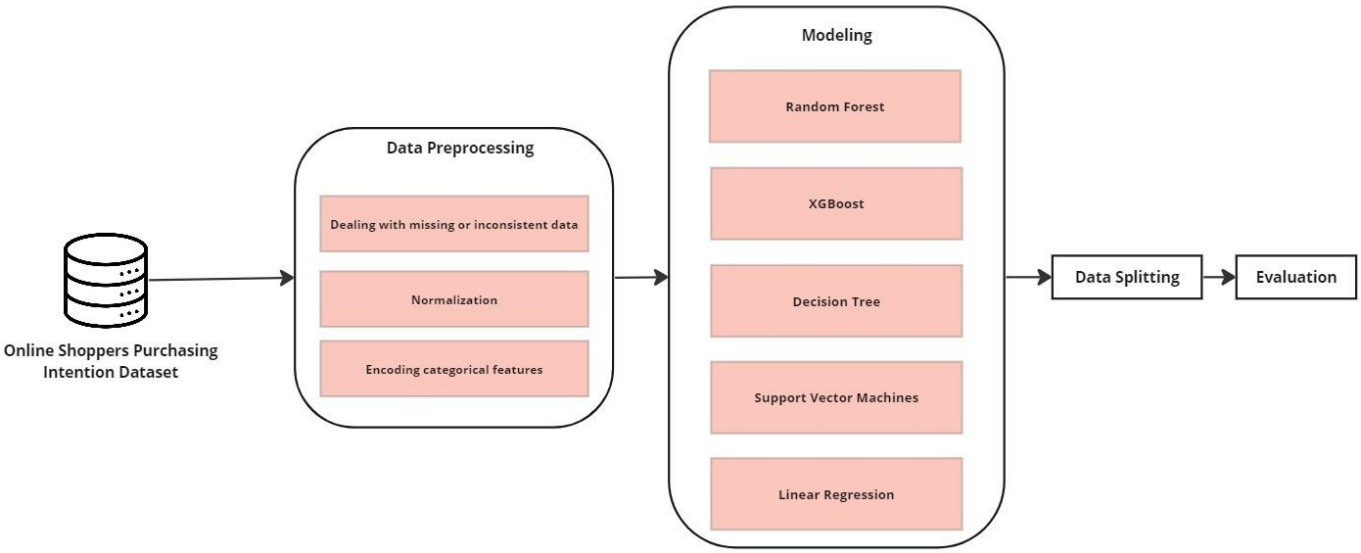

Figure Field 1: Proposed approach

As illustrated in Figure 1, in our proposed approach to predicting revenue using the Online Shoppers Purchasing Intention Dataset [14], we initially commence by procuring the dataset, which encapsulates both numerical and categorical attributes that contribute to the final purchase decision. Our second step revolves around preprocessing the data. Here, we strive to enhance the quality of our data by dealing with missing or inconsistent entries, normalizing the numerical features to ensure they're on the same scale, and encoding categorical features into a format that's comprehensible by our ML algorithms. For the third step, we transition into modeling where we employ a range of diverse algorithms such as Random Forest [15], XGBoost [16], Decision Tree [17], Support Vector Machines (SVM) [18], and Linear Regression [19]. These models cater to different data trends and complexities, giving us a comprehensive toolkit to work with, thereby improving the robustness of our approach. Next, we proceed to data splitting, where we follow a conventional 70/30 split, allocating 70% of the data for training our models and reserving the remaining 30% for testing the performance of our trained models. Our final step encompasses the evaluation of the models. We assess each model's performance using metrics appropriate for a binary classification problem, such as accuracy, precision, recall, F1-score, and area under the ROC curve. This detailed evaluation will help us in selecting the best model for revenue prediction, as well as understanding how different models perform on the task.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Bibliography

[1] OCDE (2013), « Electronic and Mobile Commerce », Documents de travail de l'OCDE sur l'économie numérique, n° 228, Éditions OCDE, Paris, https://doi.org/10.1787/5k437p2gxw6g-en.
[2] Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26* (pp. 462-474). Springer International Publishing.
[3] Li, J., Cui, T., Yang, K., Yuan, R., He, L., & Li, M. (2021). Demand Forecasting of E-Commerce Enterprises Based on Horizontal Federated Learning from the Perspective of Sustainable Development. Sustainability, 13(23), 13050. doi: 10.3390/su132313050
[4] Goldberg, J. (2022). E-commerce sales grew 50% to $870 billion during the pandemic. Forbes. Retrieved from https://www.forbes.com/sites/jasongoldberg/2022/02/18/e-commerce-sales-grew-50-to-870-billion-during-the-pandemic/?sh=727980e44e83
[5] Topic: E-commerce worldwide. (2023, July 17). Retrieved from https://www.statista.com/topics/871/online-shopping/#editorsPicks
[6] Madasoglu, P. (2023). +30 E-commerce Growth Statistics to Maximize Sales in 2023. Popupsmart. Retrieved from https://popupsmart.com/blog/ecommerce-growth-statistics
[7] Azizi, V., & Hu, G. (2019). Machine Learning Methods for Revenue Prediction in the Google Merchandise Store. Smart Service Systems, Operations Management, and Analytics. Springer. doi: 10.1007/978-3-030-30967-1_7
[8] Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. Hindawi. doi: 10.1155/2021/2179692
[9] Singh, K., Booma, P. M., & Eaganathan, U. (2020, December). E-commerce system for sale prediction using machine learning techniques. In *Journal of Physics: Conference Series* (Vol. 1712, No. 1, p. 012042). IOP Publishing.
[10] Z. Huo, "Sales Prediction based on Machine Learning," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), Hangzhou, China, 2021, pp. 410-415, doi: 10.1109/ECIT52743.2021.00093.
[11] Qi, Y., Li, C., Deng, H., Cai, M., Qi, Y., & Deng, Y. (2019). A Deep Neural Framework for Sales Forecasting in E-Commerce. CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery. doi: 10.1145/3357384.3357883
[12] Li, J., Wang, T., Chen, Z., & Luo, G. (2019). A machine learning algorithm generated sales predictions for inventory optimization in cross-border e-commerce. *International Journal of Frontiers in Engineering Technology*, *1*(1), 62-74.
[13] Dong, W., Li, Q., & Zhao, H. V. (2019). Statistical and Machine Learning-Based E-Commerce Sales Forecasting. ICCSE'19: Proceedings of the 4th International Conference on Crowd Science and Engineering. Association for Computing Machinery. doi: 10.1145/3371238.3371256
[14] Online Shoppers Purchasing Intention Dataset—UCI Machine Learning Repository. (2023, July 17). Retrieved from https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset
[15] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
[16] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). XGBoost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.
[17] Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*, 261-283.
[18] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: applications, challenges, and trends. *Neurocomputing*, *408*, 189-215.
[19] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140-147.